

Weighted Quantile Regression for Analyzing Health Care Cost Data with Missing Covariates

Ben Sherwood^a, Lan Wang^a and Xiao-Hua Zhou^{*b c}

Analysis of health care cost data is often complicated by a high level of skewness, heteroscedastic variances and the presence of missing data. Most of the existing literature on cost data analysis have been focused on modeling the conditional mean. In this paper, we study a weighted quantile regression approach for estimating the conditional quantiles health care cost data with missing covariates. The weighted quantile regression estimator is consistent, unlike the naive estimator, and asymptotically normal. Furthermore, we propose a modified BIC for variable selection in quantile regression when the covariates are missing at random. The quantile regression framework allows us to obtain a more complete picture of the effects of the covariates on the health care cost, and is naturally adapted to the skewness and heterogeneity of the cost data. The method is semiparametric in the sense that it does not require to specify the likelihood function for the random error or the covariates. The weighted quantile regression procedure and the modified BIC are investigated via extensive simulations. We illustrate the application by analyzing a real data set from a health care cost study. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Health care cost data; Missing Data; Inverse Probability Weighting; Quantile Regression.

1. Introduction

Health care cost data are characterized by a high level of skewness and heteroscedastic variances. In practice, analyzing cost data is often further complicated by the presence of missing data. When the covariates information is collected through a questionnaire or interview, non-response is a typical source for missing data; when the data are obtained from hospital records, incomplete records may lead to missing information. Missing data may also arise because the patients drop out of the study or are lost to follow up. These features pose great challenges for statistical analysis of health care cost data.

Most of the existing literature on health care cost data analysis have been focused on modeling the conditional mean (or average) of the health care cost given the covariates such as age, gender, race, marital status and disease status. The conditional mean framework has two important limitations. First, the application of the conditional mean regression model to health care cost data analysis is usually not straightforward. Due to the presence of skewness and nonconstant

^aSchool of Statistics, University of Minnesota, 313 Ford Hall, 224 Church St SE, Minneapolis, MN 55455.

^bHSR&D, VA Puget Sound Health Care System, 1100 Olive Way, 1400, Seattle, WA 98101, U.S.A.

and Department of Biostatistics, University of Washington, F600, HSB, Box # 357232, Seattle, WA 98198, U.S.A. ^c School of Statistics, Renmin University of China, Beijing, China

Contract/grant sponsor: The research is supported by National Science Foundation grant DMS-1007603.

variances, transformation of the response variable is often required when constructing the mean regression model and retransformation is needed in order to obtain direct inference on the mean cost. Second, the conditional mean model focuses primarily on the marginal effects of the risk factors on the central tendency of the conditional distribution. When the marginal effects vary across the conditional distribution, focusing on the marginal effects at the central tendency may substantially distort the information of interest at the tails. For example, a weak relationship between a risk factor and the mean health care cost does not preclude a stronger relationship at the upper or lower quantiles of the conditional distribution.

In this paper, we study an alternative approach for analyzing health care cost data based on quantile regression. The quantile regression model is a relatively new statistical tool to the field of health care cost analysis. A brief introduction to quantile regression is given in Section 2. In short, quantile regression estimates the conditional quantile function of the response variable Y given the covariates, for example the 0.5 conditional quantile or the conditional median. The knowledge of how the covariates influence high cost can be obtained by estimating a high quantile of the conditional distribution, for example the 0.9 conditional quantile. By considering different quantiles, we are able to obtain a more complete picture of the effects of the covariates on health care cost. As health care cost data often contains covariates with missing values, in this paper we investigate weighted quantile regression for parameter estimation and a new BIC criterion for variable selection in the presence of missing covariates. It is well known that statistical analysis based on observations with complete information is often biased, as those observations do not necessarily constitute a representative sample from the underlying population. Furthermore, the problem of variable selection is crucial for identifying significant risk factors that contribute to high health care cost. The knowledge of these significant risk factors is of fundamental importance for controlling the growth of health care cost.

Two popular methods for mean regression with missing data are weighting and imputation. The imputation approach imputes values for the missing data and performs the same analysis as if the data were complete, while the weighting method appropriately weight the data points with complete observations and then performs the analysis on the weighted data. The imputation approach often requires the specification of a joint or conditional likelihood. It is usually more efficient than the weighting method when the likelihood function is correctly specified. However, correct specification of the likelihood function is often challenging in practice, especially for skewed and heteroscedastic data or when the missing data contain both continuous and discrete variables. When the likelihood function is misspecified, the imputation approach may lead to biased estimation. The quantile regression based weighting approach we study in this paper is semiparametric and circumvents the difficulty of specifying the joint or conditional likelihood function. In particular, it requires no parametric distributional assumptions for either the covariates or the error term.

Research on inverse probability weighted quantile regression with missing data has been limited. Lipsitz, et. al. [1] and Yi and He [2] studied quantile regression for modeling longitudinal data with dropouts where the covariates are time invariant (thus are known at all time points) but the response variable may be missing from a certain time point. The weighted estimators in these two papers are defined by weighted estimating equations. We consider a different setting where the covariates are missing at random and study an estimator defined as the minimizer of a weighted quantile objective function. In a very recent paper, Wei et al. [3] developed a multiple imputation approach for estimating the conditional quantile in the presence of missing covariates. Their approach first imputes the missing values and then performs regular quantile regression analysis. Comparing to our proposed procedure, the multiple imputation approach requires the stronger missing completely at random (MCAR) assumption and requires a parametric model for the covariates distribution for the imputation.

The main idea of our method is *inverse probability weighting*, that is, we weight the completely observed cases inversely proportionally to the probability of being observed. The proposed methodology and theory extend those for the inverse probability weighted conditional mean regression models (e.g., [4]). We investigate the asymptotic normality of the weighted estimator and reveal an interesting phenomenon that using estimated weights often leads to asymptotically more efficient estimator than using the true weights, which echoes the same finding for weighted estimation in linear

mean regression model [4]. We also prove that a modified BIC for variable selection in quantile regression with covariates missing at random enjoys the property of variable selection consistency. The rest of the paper is organized as follows. Section 2 briefly introduces quantile regression. In Section 3, we consider quantile regression with missing covariates. When the covariates are missing at random, we define the weighted estimator and introduce a modified BIC for variable selection. In Section 4, we investigate the statistical properties of the weighted quantile estimator and the consistency of the modified BIC. Section 5 reports results from extensive Monte Carlo studies. Section 6 applies the weighted quantile regression approach to the analysis of a health care cost data set. Section 7 concludes the paper with some discussions.

2. Introduction to quantile regression

To formally define the conditional quantile function, we consider a random sample from the distribution of $\{Y, X\}$, where Y denotes the health care cost and $X = (x_0, x_1, \dots, x_d)'$ denotes the vector of covariates representing patients' characteristics. We use $'$ to denote the transpose of a vector. Following the convention, we set $x_0 = 1$. The conditional distribution function of Y given X is $F(y|X) = P(Y \leq y|X)$. For $0 < \tau < 1$, the τ th conditional quantile of Y given X is defined as $Q_{Y|X}(\tau) = \inf\{t : F(t|X) \geq \tau\}$. The case $\tau = 1/2$ corresponds to the conditional median. The 0.9 conditional quantile function is easy to interpret: given the observed vector of covariates X , 10% of all patients from the population of potential patients with the same set of covariates would incur health care cost above $Q_{Y|X}(0.9)$. A useful property of the conditional quantile function is its invariance to any monotone transformation of the response variable, that is, for any monotone function $h(\cdot)$, we have $Q_{h(Y)|X}(\tau) = h(Q_{Y|X}(\tau))$.

The linear quantile regression model assumes that $Q_{Y|X}(\tau) = X'\beta(\tau)$, where $\beta(\tau)$ denotes the unknown vector of parameters. Given n independent and identically distributed observations (Y_i, X_i) , where $X_i = (x_{i0}, x_{i1}, \dots, x_{id})'$, $i = 1, \dots, n$, we may write

$$Y_i = X_i'\beta(\tau) + u_i, \quad (1)$$

where the u_i are independent and satisfy $P(u_i < 0 | X_i) = \tau$. This is the model we will consider in the rest of the paper. Note that we do not require the error terms to be identically distributed. The regression coefficient $\beta(\tau)$ may vary across different quantiles, which is useful for modeling heterogeneity in the data. The unknown $\beta(\tau)$ can be estimated by

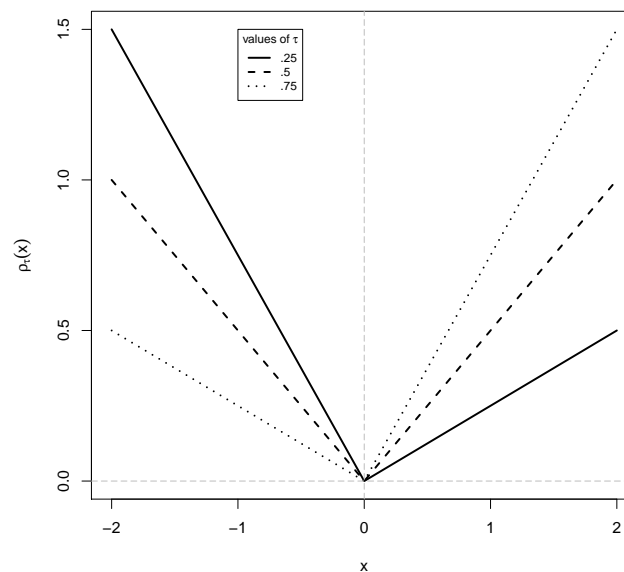
$$\hat{\beta}_n(\tau) = \operatorname{argmin}_{\beta} \sum_{i=1}^n \rho_{\tau}(Y_i - X_i'\beta), \quad (2)$$

where $\rho_{\tau}(x) = x\{\tau - I(x < 0)\}$ is the quantile loss function. The function $\rho_{\tau}(x)$ is an asymmetric L_1 loss function. Figure 1 displays the function $\rho_{\tau}(x)$ for $\tau = 0.25, 0.5$ and 0.75 , respectively. Minimization of the convex but nonsmooth objective function in (2) can be efficiently implemented using linear programming. Various statistical software packages including R, SAS, STATA, among others, provide functions for computing the quantile regression estimator. Under mild regularity conditions $\hat{\beta}_n(\tau)$ has an asymptotically normal distribution [5]. We refer to Koenker [6] for a comprehensive introduction to quantile regression.

3. Weighted quantile regression with missing covariates

Missing covariates frequently occur in health care cost data. The patients may refuse to answer certain questions, a nurse may forget to make all of the measurements or the patients may miss a follow-up appointment. Assume that we collect data on n subjects. For subject i , $i = 1, \dots, n$, we observe a response variable Y_i , a vector $W_i = (W_{i1}, \dots, W_{ip})'$ of p

Figure 1. Plot of quantile loss functions



covariates that is always fully observed, and a vector $V_i = (V_{i1}, \dots, V_{iq})'$ of q covariates that may contain some missing components. We write $X_i = (W_i', V_i')'$, the vector of all $(p + q)$ covariates. For each observation, we use an indicator variable R_i to denote if V_i is fully observed, that is, $R_i = 1$ if V_i is fully observed, and $R_i = 0$ otherwise.

We assume that V_i is missing at random (MAR) in the sense that

$$P(R_i = 1 \mid Y_i, X_i) = P(R_i = 1 \mid Y_i, W_i),$$

The MAR assumption implies that R_i and V_i are conditionally independent given Y_i and W_i . In other words, the probability of missing may depend on observed data but does not depend on the variables that are not observed. The MAR assumption is common in the missing data literature and is reasonable in many practical situations, see [7] for related discussions. We assume that $Q_{Y|X}(\tau) = X'\beta(\tau)$ for an unknown parameter $\beta(\tau)$, or equivalently model (1) holds. The goal is to estimate $\beta(\tau)$ when V_i is missing at random. Under the MAR assumption, we assume that for an unknown γ and $T_i = (Y_i, W_i')' \in \mathbb{R}^q$ we have

$$P(R_i = 1 \mid Y_i, X_i) = \pi(T_i, \gamma), \quad (3)$$

for some function $\pi(\cdot, \gamma)$, whose form is known up to a finite-dimensional parameter γ .

3.1. Weighted quantile regression

To handle the missing covariates when quantile regression is applied, a naive approach is to fit the model using only data points with complete observation. The naive estimator is

$$\hat{\beta}_n^N(\tau) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n R_i \rho_\tau(Y_i - X_i' \beta). \quad (4)$$

For linear mean regression with covariates missing at random and missingness dependent on the response, it is known that this approach often leads to a biased estimator [4]. To see that the naive estimator may be biased, we first observe that (4)

implies that the estimator $\widehat{\beta}_n^N(\tau)$ approximately solves the following estimating equation

$$G_n(\beta) = \sum_{i=1}^n R_i X_i \Psi_\tau(Y_i - X_i' \beta) = 0, \quad (5)$$

where $\Psi_\tau(t) = \tau - I(t < 0)$ is the gradient function of $\rho_\tau(t)$. From a straightforward calculation, under the covariates missing at random assumption,

$$E \left[\sum_{i=1}^n R_i X_i \Psi_\tau(Y_i - X_i' \beta) \right] = E \left[\pi(T_i, \gamma) X_i \Psi_\tau(Y_i - X_i' \beta) \right].$$

Note that $E[\Psi_\tau(Y_i - X_i' \beta(\tau)) | X_i] = 0$. However since $\pi(T_i, \gamma)$ is a function of Y_i , it is not necessarily conditionally independent of $\Psi_\tau(Y_i - X_i' \beta)$ given X_i . In general, we may not have $E[\pi(T_i, \gamma) X_i \Psi_\tau(Y_i - X_i' \beta(\tau))] = 0$, which is a necessary condition for $\widehat{\beta}_n^N(\tau)$ to be consistent for $\beta(\tau)$.

To eliminate the bias, we consider a weighted estimator based on the inverse probability weights. That is, we weight each complete record by the inverse of its probability being observed. The intuition is that under the MAR assumption for every complete observation with covariates X_i we would expect $\frac{1}{\pi(T_i, \gamma)}$ complete observations with the same covariates if there were no missing data. In the case the function $\pi(T_i, \gamma)$ is known, we estimate $\beta(\tau)$ by minimizing the weighted quantile function $\sum_{i=1}^n \frac{R_i}{\pi(T_i, \gamma)} \rho_\tau(Y_i - X_i' \beta)$. The weighted quantile regression estimator approximately solves the following weighted estimating equation

$$G_n^W(\beta) = \sum_{i=1}^n \frac{R_i}{\pi(T_i, \gamma)} X_i \Psi_\tau(Y_i - X_i' \beta) = 0. \quad (6)$$

To see that the weighted estimating equation is unbiased, we observe that by the iterative expectation formula and the MAR assumption,

$$\begin{aligned} & E \left[\frac{R_i}{\pi(T_i, \gamma)} X_i \Psi_\tau(Y_i - X_i' \beta(\tau)) \right] \\ &= E \left[E \left[\frac{R_i}{\pi(T_i, \gamma)} X_i \Psi_\tau(Y_i - X_i' \beta(\tau)) \mid X_i, Y_i \right] \right] \\ &= E \left[\frac{\pi(T_i, \gamma)}{\pi(T_i, \gamma)} X_i \Psi_\tau(Y_i - X_i' \beta(\tau)) \right] = E \left[X_i E[\Psi_\tau(Y_i - X_i' \beta(\tau)) | X_i] \right] = 0. \end{aligned}$$

In practice, the missing data mechanism is often unknown and needs to be modeled. To model the probability of observing V_i , we apply the commonly used logistic regression model for (3), which implies

$$\pi(T_i, \gamma) = \frac{e^{T_i' \gamma}}{1 + e^{T_i' \gamma}}.$$

Let $\widehat{\gamma}$ be the estimator of γ based on the logistic regression model. The weighted quantile regression estimator is formally defined as

$$\widehat{\beta}_n^W = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \frac{R_i}{\pi(T_i, \widehat{\gamma})} \rho_\tau(Y_i - X_i' \beta). \quad (7)$$

As the objective function in (7) is a weighted quantile objective function, the estimator $\widehat{\beta}_n^W$ can be easily computed using existing software. In the subsequent numerical studies, we apply the function “rq” in the R package *quantreg* with corresponding weights $\frac{R_i}{\pi(T_i, \widehat{\gamma})}$. The properties of the weighted quantile estimator $\widehat{\beta}_n^W$ will be studied in Section 4.

3.2. Modified BIC for quantile regression with missing covariates

In many practical problems, some of the covariates may be irrelevant or redundant for modeling the response variable. Including those redundant covariates in the statistical model often impairs the efficiency of estimation. It is thus important to perform variable selection before further analysis. Schwarz's BIC is a widely applied variable selection procedure. In the linear mean regression setting without missing data, it is known that under mild conditions the BIC is consistent in the sense that it selects the true model with probability approaching one if the true model is among the candidate models. When there is no missing data, BIC has been extended to quantile regression [8] and rank regression [9]. However, the use of BIC has been mostly restricted to the case the data are completely observed. Even for linear mean regression, there is no existing theory on the validity of BIC in the presence of missing data, to our best knowledge. In the following, we demonstrate that a minor modification using inverse probability weighting leads to consistent model selection for quantile regression when the covariates are missing at random.

We write $X_i = (W_i', V_i')' = (X_{i1}, \dots, X_{i(p+q)})'$. We begin by indexing each candidate model by a $(p+q)$ -dimensional binary vector $\nu = (\nu_1, \dots, \nu_{p+q})'$, where ν_j is one if the j th component of X_i belongs to the candidate model and is zero otherwise. The total number of ones in ν is denoted by d_ν , which describes the model complexity. Let $X_{i\nu}$ be the d_ν -dimensional subvector of X_i that contains the covariates in model ν ; and let β_ν be the corresponding d_ν -dimensional subvector of parameters.

In the setting of quantile regression with missing covariates, the modified BIC for the candidate model ν is defined as

$$\text{BIC}(\nu) = \min_{\beta_\nu} \left\{ \sum_{i=1}^n \frac{R_i}{\pi(T_i, \hat{\gamma})} \rho_\tau(Y_i - X_{i\nu}' \beta_\nu) + \frac{d_\nu \log n}{2} \right\}. \quad (8)$$

where $\hat{\gamma}$ is the estimator from the logistic regression model using all candidate covariates. Alternatively, we may also consider variable selection for the logistic regression model first. We select the candidate model with the smallest modified BIC value. The property of the modified BIC will be studied in Section 4.

4. Statistical properties

Under appropriate regularity conditions, the weighted quantile regression estimator $\hat{\beta}_n^W$ defined in (7) is asymptotically normal. Before we formally state this result, we introduce some notation. Let

$$\begin{aligned} D_0 &= E \left[X_i X_i' \frac{1}{\pi(T_i, \gamma)} \Psi_\tau^2(u_i) \right], \\ D_1 &= E \left[f_i(0|X_i) X_i X_i' \right], \end{aligned}$$

where $f_i(t|X_i)$ denotes the conditional density function of u_i given X_i . We also define

$$\begin{aligned} I(\gamma) &= E \left[T_i T_i' \pi(T_i, \gamma) (1 - \pi(T_i, \gamma)) \right], \\ D_2 &= E \left[(1 - \pi(T_i, \gamma)) T_i X_i' \Psi_\tau(u_i) \right]. \end{aligned}$$

We assume that the symmetric matrices D_0 , D_1 and $I(\gamma)$ are positive definite.

Theorem 4.1 Assume Conditions 1-6 in the Appendix are satisfied, then

$$\sqrt{n}(\hat{\beta}_n^W - \beta(\tau)) \rightarrow N(0, D_1^{-1} (D_0 - D_2' I(\gamma)^{-1} D_2) D_1^{-1}) \quad (9)$$

in distribution.

The proof of Theorem 4.1 is available in the Supporting Materials of this article (Online Appendix A).

Remark: For linear mean regression with covariates missing at random, an interesting phenomenon has been revealed in the literature: the inverse probability weighted estimator with the estimated weights is more efficient than the estimator using the true weights [4]. Following Theorem 4.1, we can conclude that the same phenomenon remains true for the weighted quantile regression estimator. Suppose we use the true weights $\pi(T_i, \gamma)$ instead of the estimated weights $\pi(T_i, \hat{\gamma})$ in obtaining the weighted quantile estimator and we denote this estimator by $\hat{\beta}_n^{TW}$. Following similar derivations as in the proofs for Theorem 4.1, it is easy to see that

$$\sqrt{n}(\hat{\beta}_n^{TW} - \beta(\tau)) \xrightarrow{d} N(0, D_1^{-1} D_0 D_1^{-1}).$$

Note that $D_1^{-1} (D_0 - D_2' I(\gamma)^{-1} D_2) D_1^{-1} \leq D_1^{-1} D_0 D_1^{-1}$, where for symmetric matrices A and B the notation $A \leq B$ means $t' A t \leq t' B t$ for any vector $t \neq 0$ of appropriate dimension. Hence, it is asymptotically more efficient to use the estimated weights. A heuristic explanation on why replacing $\pi(T_i, \gamma)$ by an estimate $\pi(T_i, \hat{\gamma})$ can improve the efficiency is given in Section 6.1 of Robins et al. [4].

Next, we consider the property of the modified BIC criterion. Consider a class of finitely many candidate models, each indexed by a $(p + q)$ -dimensional binary vector ν , as discussed in Section 3.2.

Theorem 4.2 Assume that this class contains the true model, which is indexed by ν_0 . Let the model selected by the modified BIC be indexed by $\hat{\nu}$, and assume that Conditions 1-7 in the Appendix are satisfied. Then as $n \rightarrow \infty$,

$$P(\hat{\nu} = \nu_0) \rightarrow 1 \quad (10)$$

Therefore, the modified BIC for quantile regression with covariates missing at random possesses the property of model selection consistency. The proof of this result is available in the Supporting Materials of this article (Online Appendix B).

5. Simulations

5.1. Simulations for the weighted quantile estimator

We now investigate the performance of the weighted quantile regression estimator through Monte-Carlo studies. We generate random data from the following model:

$$Y_i = -3 + X_{i1} - X_{i2} + X_{i3} + \sigma \epsilon_i, \quad i = 1, \dots, 100,$$

where $X_{i1} \sim N(0, 1)$, $X_{i2} \sim N(0, 1)$ and $X_{i3} \sim \text{Bernoulli}(0.5)$ are independent. In our simulations, X_{i2} is always observed while (X_{i1}, X_{i3}) is missing at random. We consider three different distributions for the random error ϵ_i : (1) standard normal distribution; (2) heteroscedastic normal distribution $\epsilon_i = (1 + X_{i3})Z_i$ where $Z_i \sim N(0, 1)$; and (3) mixture normal distribution with 90% of the data from $N(0, 1)$ and 10% from $N(0, 25)$.

Let R_i be a binary variable indicating if (X_{i1}, X_{i3}) is fully observed. We generate the R_i from the following logistic regression model

$$\text{logit}(P(R_i = 1 \mid X_i, Y_i)) = 4 + Y_i + X_{i2},$$

where $\text{logit}(t) = \log(\frac{t}{1-t})$. Hence, the MAR assumption for (X_{i1}, X_{i3}) is satisfied. The above missing data mechanism produces on average a missing rate ranging from 24% to 30% for the different distributions of ϵ_i .

Table 1 summarizes the results for estimating the regression coefficients in $Q_{Y_i|X_{i1}, X_{i2}, X_{i3}}(\tau) = \beta_0(\tau) + \beta_1(\tau)X_{i1} + \beta_2(\tau)X_{i2} + \beta_3(\tau)X_{i3}$, for $\tau = 0.5$ and 0.8 , when $\sigma = 1$. We compare the proposed weighted quantile regression procedure (called “weighted quantile”) with the naive quantile method defined in (4) (called “naive quantile”), and the oracle procedure (called “oracle”) which works by making use of the knowledge of the missing data and thus is not implementable in real data analysis. For $\tau = 0.5$, we also compare with the weighted mean regression estimator (called “weighted mean”), which is the weighted linear mean estimator using the inverse probability weighting approach. The results for these four procedures are summarized based on $M = 500$ simulation runs. For each τ , we report the bias and standard error for estimating each of the four coefficients along with the overall mean squared error (MSE), where $\text{MSE} = M^{-1} \sum_{k=1}^M \sum_{j=0}^3 [\hat{\beta}_j^k(\tau) - \beta_j(\tau)]^2$, with $\hat{\beta}_j^k(\tau)$ being the estimator for $\beta_j(\tau)$ for the k th simulation run.

Put Table 1 about here

From Table 1, we observe that: (1) the naive quantile regression estimator is biased; (2) the weighted quantile estimator performs satisfactorily comparing with the oracle estimator; (3) for $\tau = 0.5$, the weighted mean method works well if the random error is normally distributed but is less efficient when the error is heavy-tailed. We have also performed the same simulation study when $\sigma = 0.5$, for which similar phenomena were observed but are not reported here due to limited space. We note that the bias of the naive quantile regression estimator is more pronounced for the larger value of σ .

Table 1. Simulation results for Estimation

Error Distribution	Method	Bias ($\hat{\beta}_0$)	Bias ($\hat{\beta}_1$)	Bias ($\hat{\beta}_2$)	Bias ($\hat{\beta}_3$)	SE ($\hat{\beta}_0$)	SE ($\hat{\beta}_1$)	SE ($\hat{\beta}_2$)	SE ($\hat{\beta}_3$)	MSE	τ
N(0,1)	weighted mean	0.02	-0.03	0.00	-0.01	0.20	0.14	0.13	0.28	0.16	0.50
N(0,1)	oracle	-0.01	0.00	0.00	-0.00	0.18	0.13	0.13	0.26	0.13	0.50
N(0,1)	naive quantile	0.24	-0.10	0.00	-0.10	0.21	0.15	0.14	0.29	0.25	0.50
N(0,1)	weighted quantile	-0.01	-0.00	-0.00	-0.01	0.24	0.18	0.17	0.34	0.23	0.50
Heteroscedastic	weighted mean	0.01	-0.06	-0.01	0.10	0.21	0.26	0.24	0.51	0.44	0.50
Heteroscedastic	oracle	-0.01	-0.00	0.01	-0.01	0.17	0.16	0.16	0.39	0.24	0.50
Heteroscedastic	naive quantile	0.28	-0.19	-0.01	0.26	0.21	0.21	0.18	0.41	0.47	0.50
Heteroscedastic	weighted quantile	-0.03	-0.01	-0.00	0.04	0.26	0.29	0.28	0.61	0.60	0.50
Mixed Normal	weighted mean	0.16	-0.05	-0.01	-0.05	0.33	0.25	0.25	0.53	0.54	0.50
Mixed Normal	oracle	-0.01	0.00	0.00	0.01	0.20	0.14	0.14	0.29	0.16	0.50
Mixed Normal	naive quantile	0.32	-0.12	-0.00	-0.13	0.24	0.17	0.16	0.33	0.36	0.50
Mixed Normal	weighted quantile	-0.02	-0.01	0.00	0.00	0.32	0.25	0.25	0.53	0.50	0.50
N(0,1)	oracle	-0.04	-0.00	0.01	0.02	0.21	0.15	0.14	0.28	0.17	0.80
N(0,1)	naive quantile	0.17	-0.09	0.00	-0.08	0.25	0.17	0.16	0.33	0.27	0.80
N(0,1)	weighted quantile	-0.04	0.01	0.01	0.02	0.24	0.19	0.16	0.32	0.23	0.80
Heteroscedastic	oracle	-0.00	-0.01	-0.01	-0.02	0.21	0.19	0.20	0.45	0.32	0.80
Heteroscedastic	naive quantile	0.20	-0.16	-0.01	0.14	0.26	0.23	0.23	0.50	0.51	0.80
Heteroscedastic	weighted quantile	-0.01	0.00	-0.02	0.02	0.25	0.27	0.26	0.56	0.52	0.80
Mixed Normal	oracle	-0.01	-0.01	-0.01	-0.01	0.26	0.18	0.18	0.35	0.25	0.80
Mixed Normal	naive quantile	0.31	-0.13	-0.01	-0.15	0.32	0.23	0.21	0.41	0.49	0.80
Mixed Normal	weighted quantile	0.01	-0.01	-0.01	-0.00	0.31	0.22	0.22	0.45	0.39	0.80

5.2. Simulations for variable selection

Next, we examine the performance of the modified BIC for quantile regression with covariates missing at random. We consider a similar model as in Section 5.1 for generating the random response variable

$$Y_i = -3 + X_{i1} - X_{i2} + X_{i3} + \sigma\epsilon_i, \quad i = 1, \dots, 100, \quad (11)$$

where X_{i2} is always observed while (X_{i1}, X_{i3}) is missing at random. To demonstrate variable selection, we add three redundant covariates (X_{i4}, X_{i5}, X_{i6}) , which are independent and identically distributed $N(0, 1)$ random variables and are independent of the predictors and error term in (11). The true conditional quantile function of Y_i thus contains (X_{i1}, X_{i2}, X_{i3}) . We consider the same three different distributions for ϵ_i and report results only for values of $\sigma = 1$. In the unreported results, we also observe that the modified BIC performing better for $\sigma = 0.5$. The missingness indicator R_i is generated from the following logistic regression model

$$\text{logit}(P(R_i = 1 \mid X_i, Y_i)) = 4 + Y_i + X_{2i} + X_{4i} + X_{5i} + X_{6i}.$$

The missing rates in all the simulation settings on average are around 30%. In the simulations, we estimate the missing probability by fitting the above logistic regression model with all six candidate covariates included.

Table 2 summarizes the simulation results based on 500 simulation runs for selecting the models at $\tau = 0.5$ and 0.8, respectively. We compare the modified BIC for quantile regression (called “weighted quantile”) with the omniscient procedure (called “omniscient”) which applies the BIC while assuming the complete knowledge of all observations. A model is said to be a “correct” fit if it selects exactly the covariates $\{X_{i1}, X_{i2}, X_{i3}\}$; it is said to be an “underfit” if it misses one or more covariates from $\{X_{i1}, X_{i2}, X_{i3}\}$; and it is said to be an “overfit” if it selects $\{X_{i1}, X_{i2}, X_{i3}\}$ but at the same time includes one or more covariates from $\{X_{i4}, X_{i5}, X_{i6}\}$. We record the percentage of times the selected model is “correct”, “underfit” and “overfit”. We also record the MSE of the finally selected model for each of the two procedures. The modified BIC for quantile regression performs satisfactorily in terms of the probability of selecting the true model and estimating the coefficients.

Put Table 2 about here

6. Application to analysis of health care cost data

The data we analyze came from a clinical study on the cost-effectiveness of a computer-assisted prospective drug utilization review program conducted by Tierney et al [10]. The study was conducted in the primary care system of Indiana University Medical Group Primary Care. The data set was analyzed by Zhou, Stroupe, Tierney [11] using a heteroscedastic mean regression model. In their analysis, patients with missing information have been excluded. In this new analysis, we also include patients with missing information and focus on quantile profile of patient’s costs, which can help us to identify high-cost patients. The response variable “charge” is the log-transformed amount (\$) charged for the health care on each of the 695 patients. We consider seven covariates, which were identified in Zhou et al. to be important for modeling the cost of health care. The seven covariates are: aa (a binary variable indicating whether the patient is African-American), female (a binary variable indicating whether the patient is female), pharm_sat (pharmacist satisfaction score), alone (a binary variable indicating whether the patient is living alone), SF36_PF (SF-36 physical function score), badReaction (a binary variable indicating whether the patient stops medication because of adverse effects) and sexuallyActive (a binary variable indicating whether the patient engages in sexual activity). The pharmacist satisfaction score is based on the rating by the

patient of their satisfaction of the pharmacist (on a scale 1-5) with an average score of 2.35; while the SF-36 physical function score is based on the rating of the patients physical fitness (on a scale 1-100) with an average score of 49.6. About 10% patients have missing values on the covariates vector (pharm_sat, SF36.PF), while all the other covariates are fully observed for all the patients.

Here, we demonstrate that the weighted quantile regression approach provides new perspectives on this data set. With a small percentage of patients accounting for most of the health care costs, it is of particular interest to consider the patients with high costs, in other words, the high conditional quantiles, such as $\tau = 0.8$ and 0.9 . Quantile regression is natural for handling the heterogeneity in the data. Due to the invariance property of quantile regression, the interpretation of quantile regression based on log transformed cost is straightforward. Finally, when the BIC procedure is applied to the high quantiles, it would help identify the potential risk factors associated with high health care costs.

We first model the missing data mechanism by fitting a logistic regression model using the missing data indicator as the response variable and using the response variable and the seven covariates as predictors. We apply a stepwise logistic regression model selection using BIC. It is found that the best logistic regression model is a simple one that contains the intercept and log(charge). Using the estimated weights, we fit the weighted quantile regression model at $\tau = 0.5, 0.8$ and 0.9 . All programming is done in R and the quantreg package was used for fitting the weighted quantile regression model. Table 3 reports the estimates of the coefficients (with the corresponding p-values given in parentheses) from both the weighted quantile regression procedure and the naive quantile regression procedure. P-values are obtained by applying bootstrap with 10,000 replications to obtain standard error estimates for the coefficients. For the naive quantile estimator, the bootstrap is performed on the data points with complete records. If “m” is the quantile regression model of interest, then the bootstrapped standard errors can be obtained by the command `summary(m, se = “boot”, R = 10000)` in R. For the weighted quantile regression estimator, we also reestimate the weights for each bootstrap sample. We observe that the estimated coefficients vary noticeably across quantiles. We also observe that the role of pharmacist satisfaction is important at the median but not as important at the upper quantiles. These observations indicate the heterogeneity in the data. Furthermore, the weighted method and naive method also give different information regarding the importance of different variables. For example, the two covariates badReaction and sexuallyActive are both considered as significant by the weighted method, but not so by the naive method at the median.

Put Table 3 about here

We then apply the modified BIC for variable selection at the three quantile levels: $\tau = 0.5, 0.8$ and 0.9 . We consider the seven candidate covariates. Table 4 summarizes the selected model (including the estimated coefficients) at each of three quantiles. We note that different subsets of covariates are selected at the three quantiles. The covariate SF36.PF, which describes the patient’s overall fitness level, is selected at all levels. It is also the only covariate selected at the median. The variable badReaction is selected at $\tau = 0.8$, and the variable female is selected at $\tau = 0.9$. Hence, different covariates are considered as important for modeling the cost at different quantiles, which again indicate the heterogeneity of the data.

Put Table 4 about here

The models selected at the three quantiles are sparse. We examine the predictive performance of the selected models by comparing with the full model. We consider a random partition of the data set: 595 patients are randomly selected for training data and the remaining 100 patients are used as testing data. We apply the modified BIC and fit the selected model using the training data for $\tau = 0.5, 0.8$ and 0.9 ; we also fit the full model to the training data at these three quantiles. Let n_i be the number of records with complete data from the testing data. Then we apply the selected model and the full model to those data points with complete records in the testing data, and evaluate their predictive performance by calculating the mean absolute prediction error $n_i^{-1} \sum_{j=1}^{n_i} |\hat{Y}_j - Y_j|$, where \hat{Y}_j is the predicted value for the j th patient.

We repeat the above random partition 500 times and report the overall mean absolute prediction error when the selected model (from modified BIC) and the full model are used, respectively. The results are summarized in Table 5. We observe that the selected sparse models have similar predictive performance comparing to the full model. Hence the modified BIC procedure effectively reduces the model complexity without sacrificing the predictive ability.

Put Table 5 about here

7. Conclusions and discussions

In the current literature of health care cost data analysis, little attention has been devoted to the important problems of handling missing data and variable selection. In this paper, we study a quantile regression framework for analyzing health care cost data with covariates missing at random. The quantile regression approach is particularly suitable for analyzing the highly skewed and heteroscedastic cost data. We consider a weighted quantile regression estimator when the covariates are missing at random and proposed a modified BIC for variable selection. Extensive numerical studies demonstrate the effectiveness of the quantile regression procedure for health care cost data analysis.

A practically important problem for applying the proposed weighted quantile regression procedure is the specification of the missing data mechanism model. In Section 5 we assume all fully observed variables are in the missing data mechanism model, while in Section 6 we apply stepwise BIC to choose the relevant covariates in the missing data mechanism model. A misspecified model may lead to potential bias. To illustrate the effect of misspecification and model selection, we consider a simulation scenario similar as that in Section 5.1 with $\tau = 0.5$ and standard error normal distribution. A slight difference is that X_1 is fully observed and has no effect on the missing rate. In Table 6, we summarize the bias, standard error and MSE of the weighted quantile regression procedure when four different working models are used for estimating the missing probability: “correct” uses a correctly specified model, “overfit” includes X_1 in the model, “underfit no y ” does not include Y , “underfit no x_2 ” does not include X_2 , and “stepwise BIC” uses backward stepwise BIC selection to identify the model.

Put Table 6 about here

From Table 6 we observe: (1) over-fitting the model has negligible bias for estimating the quantile regression coefficients; (2) misspecified model may result in significant bias particularly when Y is not included; (3) the BIC procedure produces similar results as using a saturated model. Generally speaking, careful modeling of $\pi(T_i, \gamma)$ is needed. We suggest always including Y in the model $\pi(T_i, \gamma)$. Another good strategy is to potentially overfit the logistic regression model by including all covariates because extra covariates in the model do not seem to significantly increase the standard errors for $\hat{\beta}(\tau)$ but underfitting can result in bias. An interesting future research direction is to investigate semiparametric or nonparametric modeling of the missing data model, which can lead to a more robust procedure.

In the real data analysis, following the work of Zhou, Stroupe, Tierney (2001) we analyzed the data without zero costs. However, when the interest is focused on the conditional quantile, the presence of zero cost observations does not usually cause the same complexity as for estimating the conditional mean. It is well known that the mean of a random variable can be adversely affected by the tail behavior of its distribution, but any alteration of the tail would have no impact on the median. For quantile regression, if we want to estimate the conditional median and modify the values of responses at the left tail to zero, it will not change the estimation of the conditional median (similarly for high quantiles).

The problem of model checking is an important issue. To the best of our knowledge, model lack-of-fit tests have not yet been developed for quantile regression with missing data in the literature. We believe that some of the results for model

checking for quantile regression, such as He and Zhu [13] and Wang [14], can be extended to the missing data setting using the weighted method. The weighted quantile regression estimator only uses those observations that are fully observed. The procedure is most suitable when the missing data rate is not overly high. When different covariates are missing for different covariates, it is desirable to try to make use of the partial information. There are two possible ways to extend the current procedure. One is to use augmented estimating equations (e.g., [15]) to make use of the partial information; the other is to hybrid with an imputation procedure (e.g., [3]) to fill in some of the missing values. These will be interesting future research directions.

Appendix: technical conditions for the results in Section 4

Condition 1. The random error u_i in (1) has a continuous differentiable conditional density function $f_i(\cdot|X_i)$, whose derivative $f'_i(\cdot|X_i)$ exists in an open set around zero and is uniformly bounded away from 0 and ∞ around 0 in i . Furthermore, $\max_{1 \leq i \leq n} E(u_i^4) < \infty$ and the matrix $E[f_i(0|X_i)X_iX_i'] = D_1$ is positive definite.

Condition 2. There exists a compact set G such that $X_i \in G \in \mathbb{R}^p \forall i$.

Condition 3. The matrix $E\left[X_iX_i' \frac{1}{\pi(T_i, \gamma)} \Psi_\tau(u_i)^2\right] = D_0$ is positive definite.

Condition 4. The MLE $\hat{\gamma}$ for γ satisfy the regularity conditions of asymptotic normality of MLEs for exponential family models (e.g., the conditions in Section 10.6.2 of [12]).

Condition 5. There exists $\alpha > 0$ such that $\pi(T_i, \gamma) > \alpha$ uniformly in i .

Condition 6. The matrix $I(\gamma) = E[T_iT_i'\pi(T_i, \gamma)(1 - \pi(T_i, \gamma))]$ is positive definite.

Condition 7. Let β denote the true value of the parameter for the underlying model. If β^* is the limiting value for the estimator for an incorrect model, then $E[\rho_\tau(Y_i - X_i'\beta)] < E[\rho_\tau(Y_i - X_i'\beta^*)]$.

Acknowledgement

The authors are grateful for the comments from the AE and the two referees, which help us significantly improve the paper. Wang and Zhou are joint correspondence authors for this paper. The work of Sherwood and Wang was supported by National Science Foundation grant DMS-1007603. Xiao-Hua Zhou, Ph.D., is presently a Core Investigator and Biostatistics Unit Director at the Northwest HSR&D Center of Excellence, Department of Veterans Affairs Medical Center, Seattle, WA. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs. Dr. Zhou's work was supported in part by U.S. Department of Veterans Affairs, Veterans Affairs Health Administration, HSR&D grants (RCS 05-196).

References

1. Lipsitz S, Fitzmaurice G, Molenberghs G, Zhao LP. Quantile regression methods for longitudinal data with drop-outs: application to CD4 cell counts of patients infected with the human immunodeficiency virus. *Applied Statistics* 1997; **46(4)**:463-476. DOI: 10.1111/1467-9876.00084.
2. Yi G, He W. Median Regression Models for Longitudinal Data with Dropouts. *Biometrics* 2009; **65(2)**:618-625. DOI: 10.1111/j.1541-0420.2008.01105.x.
3. Wei Y, Ma Y, Carroll RJ. Multiple imputation in quantile regression. *Biometrika* 2012; **99(2)**:423-438. DOI: 10.1093/biomet/asn007.
4. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89(427)**:846-866. DOI: 10.2307/2290910.
5. Koenker R, Bassett GW. Regression quantiles. *Econometrica* 1978; **46(1)**:33-50. DOI: 10.2307/1913643.
6. Koenker R. *Quantile Regression*. Cambridge University Press: Cambridge, 2005. DOI:10.1017/CBO9780511754098.
7. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.
8. Machado, JAF. Robust Model Selection and M-Estimation. *Econometric Theory* 1993; **9(3)**:478-493. DOI: 10.1017/S0266466600007775.
9. Wang L. Wilcoxon-type generalized Bayesian information criterion. *Biometrika* 2009; **96(1)**:163-173. DOI: 10.1093/biomet/asn060.

10. Tierney W, Fitzgerald J, Miller M, Zhou X, Harris L, Wolinsky F. Predicting inpatient costs with admitting clinical data. *Medical Care* 1995; **33**(1):1-14. DOI: 10.1097/00005650-199501000-00001.
11. Zhou X, Stroupe K, Tierney W. Regression analysis of health care charges with heteroscedasticity. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 2001; **50**(3):303-312. DOI: 10.1111/1467-9876.00235.
12. Casella G, Berger R. *Statistical Inference*. Duxbury Press: Pacific Grove, 2002.
13. He X, Zhu LX. A lack-of-fit test for quantile regression. *Journal of the American Statistical Association* 2003; **98**:1013-1022. DOI: 10.1198/016214503000000963.
14. Wang L. Nonparametric test for checking lack-of-fit of quantile regression model under random censoring. *Canadian Journal of Statistics* 2008; **36**(2):321-336. DOI: 10.1002/cjs.5550360209.
15. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* 1995; **90**:122-129. DOI: 10.1080/01621459.1995.10476494.

Table 2. Simulation results for variable selection

Error Distribution	τ	Method	Correct	Overfit	Underfit	MSE
N(0,1)	.5	omniscient	0.89	0.01	0.10	0.23
N(0,1)	.5	weighted quantile	0.57	0.19	0.24	0.59
Heteroscedastic	.5	omniscient	0.64	0.04	0.33	0.53
Heteroscedastic	.5	weighted quantile	0.43	0.23	0.34	1.03
Mixed Normal	.5	omniscient	0.84	0.01	0.14	0.31
Mixed Normal	.5	weighted quantile	0.58	0.17	0.25	0.75
N(0,1)	.8	omniscient	0.68	0.00	0.32	0.54
N(0,1)	.8	weighted quantile	0.60	0.05	0.35	0.70
Heteroscedastic	.8	omniscient	0.92	0.03	0.05	0.45
Heteroscedastic	.8	weighted quantile	0.74	0.15	0.11	0.99
Mixed Normal	.8	omniscient	0.56	0.01	0.42	0.76
Mixed Normal	.8	weighted quantile	0.46	0.09	0.45	1.07

Table 3. Analysis of health care costs data: estimation for the full model (with p-values in the parentheses)

	Weighted Median	Naive Median	Weighted 0.80	Naive 0.80	Weighted 0.90	Naive 0.90
Intercept	8.84 (0)	8.77 (0)	9.4 (0)	9.56 (0)	10.77 (0)	11.1 (0)
aa	-0.21 (0.03)	-0.19 (0.06)	-0.37 (0.03)	-0.34 (0.06)	-0.48 (0.05)	-0.38 (0.1)
female	-0.22 (0.1)	-0.26 (0.06)	-0.33 (0.11)	-0.45 (0.03)	-0.81 (0.02)	-1.05 (0)
pharm_sat	-0.2 (0.01)	-0.18 (0.03)	0 (0.99)	0.02 (0.84)	-0.06 (0.65)	-0.11 (0.42)
alone	0.1 (0.45)	0.13 (0.32)	0.52 (0.01)	0.47 (0.01)	0.47 (0.08)	0.37 (0.16)
SF36.PF	-0.01 (0)	-0.01 (0)	-0.01 (0)	-0.01 (0)	-0.01 (0)	-0.01 (0)
badReaction	0.39 (0.04)	0.36 (0.06)	0.69 (0.01)	0.64 (0.02)	0.6 (0.05)	0.66 (0.04)
sexuallyActive	-0.21 (0.05)	-0.17 (0.12)	-0.2 (0.32)	-0.21 (0.29)	-0.23 (0.36)	-0.14 (0.57)

Table 4. Analysis of health care costs data: results from variable selection using the modified BIC at $\tau = 0.5, 0.8$ and 0.9

τ	0.5	0.80	0.90
Intercept	8.09	8.99	10.68
aa	-	-	-
female	-	-	-0.91
pharm_sat	-	-	-
alone	-	-	-
SF36.PF	-0.01	-0.01	-0.01
badReaction	-	0.77	-
sexuallyActive	-	-	-

Table 5. Analysis of health care costs data: absolute prediction error (APE) of the selected model and the full model (based on 500 random partitions)

Model	τ	APE
Selected	0.50	0.96
Full	0.50	0.95
Selected	0.80	1.25
Full	0.80	1.25
Selected	0.90	1.77
Full	0.90	1.71

Table 6. Possible misspecification of the missing data model ($\tau = 0.5, N(0, 1)$ error)

Method	Bias ($\hat{\beta}_0$)	Bias ($\hat{\beta}_1$)	Bias ($\hat{\beta}_2$)	Bias ($\hat{\beta}_3$)	SE ($\hat{\beta}_0$)	SE ($\hat{\beta}_1$)	SE ($\hat{\beta}_2$)	SE ($\hat{\beta}_3$)	MSE
correct	0.01	-0.01	0.00	0.01	0.25	0.18	0.19	0.35	0.25
underfit no x_2	-0.00	-0.01	-0.00	0.01	0.26	0.18	0.19	0.36	0.26
underfit no y	0.27	-0.10	0.10	-0.10	0.22	0.15	0.16	0.30	0.28
overfit	0.00	-0.01	0.00	0.01	0.25	0.18	0.19	0.35	0.25
stepwise BIC	-0.00	-0.00	0.00	0.01	0.26	0.18	0.19	0.36	0.26